An Evaluation of AIRES and STATISTICA Text Mining Tools as Applied to General Aviation Accidents

June 2013

DOT/FAA/TC-TN13/7

This document is available to the U.S. public through the National Technical Information Services (NTIS), Springfield, Virginia 22161.

This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The U.S. Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency. This document does not constitute FAA policy. Consult the FAA sponsoring organization listed on the Technical Documentation page as to its use.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

1. Report No. DOT/FAA/TC-TN13/7	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

AN EVALUATION OF AIRES AND STATISTICA TEXT MINING TOOLS AS APPLIED TO GENERAL AVIATION ACCIDENTS

5. Report Date

June 2013

6. Performing Organization Code

7. Author(s)

Massoud Bazargan, Matthew Johnson, and Akhila Vijayanarayanan

8. Performing Organization Report No.

9. Performing Organization Name and Address

Embry–Riddle Aeronautical University
600 S. Clyde–Morris Blvd.,
Daytona Beach, FL 32114

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

07-C-GA-ERAU-028

12. Sponsoring Agency Name and Address

U.S. Department of Transportation
Federal Aviation Administration
Orville Wright Bldg (FOB10A)
FAA National Headquarters
800 Independence Ave., SW
Washington, DC 20591

13. Type of Report and Period Covered

Technical Note

14. Sponsoring Agency Code

AVP-200

15. Supplementary Notes

The Federal Aviation Administration William J. Hughes Technical Center Aviation Research Division Technical Monitor was Huasheng Li.

16. Abstract

The text mining techniques and software are examined in this report in the application to identify patterns leading to general aviation (GA) accidents. Two examples of text mining software— ASIAS Information Retrieval and Extraction System (AIRES) developed by MITRE, and STATISTICA, a commercially based software—were applied to the analysis of National Transportation Safety Board accident reports. Various analyses were conducted in both the nation and each region to identify the patterns among GA accidents. It was indicated that AIRES performs relatively better than STATISTICA in terms of predicting the patterns and associations between fatal and nonfatal accidents; however, they are both unsuccessful in generating strong relationships. They also fail to identify many patterns and relationships that can be discovered through statistical analyses.
The results of text mining are in general concurrence with the results of the logistic regression performed in the previous report. However, compared with logistic analysis, text mining is more suited for exploratory and confirmatory analyses.

17. Key Words

Text mining, General aviation, Accidents, Contributing factors, Accident pattern, Federal Aviation Administration region

18. Distribution Statement

This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 56	22. Price

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AIRES	ASIAS Information Retrieval and Extraction System
ASIAS	Aviation Safety Information Analysis and Sharing
ERAU	Embry-Riddle Aeronautical University
FAA	Federal Aviation Administration
GA	General aviation
IFR/IFC	Instrument Flight Rules/Conditions
IMC	Instrument meteorological conditions
NTSB	National Transportation Safety Board
PCA	Principal Components Analysis
SOM	Self-Organizing Maps
VFR	Visual Flight Rules

EXECUTIVE SUMMARY

The research team at Embry–Riddle Aeronautical University conducted extensive statistical analyses over the previous years to identify patterns and associations among fatal and nonfatal general aviation (GA) accidents. Using various fields for these analyses, the National Transportation Safety Board (NTSB) database was utilized. The NTSB aviation accident database also includes narrative reports by accident investigators. Therefore, it was of interest to conduct text mining analyses on these narratives to see if new patterns or associations among GA accidents could be discovered. Text mining is the process of discovering new information by analyzing data to look for patterns, trends, and relationships that are not recognized by traditional statistical techniques. Text mining involves linguistic and machine-learning techniques that model and structure text-based data for a variety of purposes. The method has been extensively employed in such fields as market research and national security/intelligence. Two examples of text mining software—Aviation Safety Information Analysis and Sharing (ASIAS) Information Retrieval and Extraction System (AIRES) developed by MITRE, and STATISTICA, a commercially based software—were used for this study. Various analyses were conducted for national and Federal Aviation Administration regions to find patterns among GA accidents using the two software. While AIRES performed relatively better than STATISTICA in terms of predicting the patterns and associations between fatal and nonfatal accidents, both were unsuccessful in generating strong relationships. The relationships that were discovered through statistical analyses were much stronger and robust when compared to text mining software. The text mining software failed to identify many patterns and relationships that were discovered through statistical analyses. These software did not generate any new reports that were not identified in our statistical analyses; one reason for this could be the fact that the narrative reports in the NTSB database do not follow a rigid format and were compiled by many investigators who used different words, terminologies, and phrases to describe the accidents.

The results of text mining are in general concurrence with the results of the logistic regression performed in the previous report. However, compared with logistic analysis, text mining is more suited for exploratory and confirmatory analyses.

INTRODUCTION

BACKGROUND.

The National Transportation Safety Board (NTSB) reports that general aviation (GA) is responsible for 82% of total air transport-related accidents and incidents and for 83% of all air transport-related fatalities. Previous work in this area includes statistical analysis in the form of logistic regression on a large sample of GA accidents from 1982 to 2009. The purpose of this analysis was to determine what factors contribute to the seriousness of injury of the involved parties in an accident. The inputs to the regression included a variety of variables available in the NTSB Aviation Accident Database, including the pilot's experience, wind and light conditions, flight phase, and aircraft characteristics. The results of this analysis indicated that several of the selected factors were statistically significant in predicting fatal GA accidents, including flying at night; performing a cross-country flight; the descent phase of flight; flying in instrument conditions; total flight hours of between 50 and 300 hours; flying while tired; flying with a second pilot; and flying with retractable gear. Some of the results of the analysis are in line with intuitive expectation, but others are counterintuitive. In particular, although intuition suggests that the presence of a second pilot would reduce the likelihood of an accident being fatal, the analysis revealed that this is not the case—perhaps because of interference with the primary pilot.

PURPOSE.

Among the primary limitations of logistic regression is the requirement that the analyst choose the independent variables for inclusion in the model. Given that a second pilot increases the likelihood of accident fatality, it is evident that the contributing factors to a fatal accident are not straightforward. Even if one were to analyze every variable collected by the NTSB and test for significance, the results of the analysis rely on the inclusiveness of the database items themselves. The NTSB aviation accident database also includes narrative reports by accident investigators. These reports do not follow a rigid format and, thus, may contain additional information not included in the structured parts of the database. This report catalogs efforts to assess the completeness and accuracy of the previous research and to provide additional insight into the nature of the sampled accidents. This is accomplished by mining the unstructured text portion of the accident database for statistical relationships.

SCOPE OF REPORT.

The research team at Embry-Riddle Aeronautical University (ERAU) mined the text of the narrative sections of the NTSB accident database and performed text analysis on the results. This report builds on previous research, including two publications (Bazargan and Guzhva, 2007 and Bazargan and Guzhva, 2011) and a previous Federal Aviation Administration (FAA) report submission (Bazargan et al., 2012). The publications describe studies conducted on a national basis, while the previous report submission is categorized regionally. The analysis presented in this report is organized on the same regional basis as the previous FAA report submission.

THE NTSB DATABASE.

The federal regulations require a pilot/operator of an aircraft to immediately notify the regional office of the NTSB nearest to the accident. An accident is defined as an occurrence during an aircraft operation that takes place between the time any person boards the aircraft with the intention of flight and all such persons have disembarked, and in which any person suffers death or serious injury, or the aircraft receives substantial damage. The NTSB uses a factual report (NTSB form 6120), which contains more than 400 fields of data pertaining to GA accidents. These reports are maintained in a publicly accessible database containing more than 60,000 aviation accidents and incidents, with more than 400 fields describing all information relevant to the accident or incident.

The FAA has provided a Microsoft® Access® file of the aircraft accident database from their Aviation Safety Information Analysis and Sharing (ASIAS) system. The database contains 66,633 unique events that took place from 1982 to 2009. Only accident data for flights under 14 Code of Federal Regulations Part 91 GA were considered in the analyses presented in this report. Only four of the more than 400 data fields were considered in the initial analysis. Of those four, the three report narratives included for each accident—preliminary, final, and cause—were mined for word clusters to include as independent variables. The injury description field was used as a dependent variable in order to examine the presence of words in the narratives for impacts on the seriousness of the injuries. Many ASIAS users have identified problems relating to the quality of the structured fields. A chief complaint from these users has been that certain elements present in the unstructured text reports that provide significant value to analysis are omitted from the structured fields. In some cases, this appears to be due to input errors (where a field exists for the relevant data and is left blank) and, in other cases, it is the result of there being no field for the relevant data. This creates challenges for traditional analysis of the database and means that many reports used to be discarded so that a data set without blanks can be created.

OBJECTIVES

PROBLEM STATEMENT.

Considering the current volume of GA accidents in the U.S., what are the primary contributing factors to the injury seriousness of these accidents, and how can these factors be mitigated? The ultimate goal of this analysis is to provide evidence that these contributory factors can be identified from unstructured text data.

RESEARCH QUESTIONS.

The research strategy for this report is to review techniques currently employed in nonaviation fields to analyze unstructured text. The review focuses on applications in safety critical and transportation fields in particular. At the onset of the project, five research questions were formulated to define the scope of the research. To explore the approaches used for text mining, the following two questions were asked:

1. What approaches are currently being adopted to obtain conclusions about situations and causal relationships based on unstructured text?

2. How do these approaches differ from traditional regression analyses and from each other? Which approaches are more appropriate responses to the problem statement?

To identify the state of the text mining and analytics field and the current view of techniques therein, two additional questions needed to be addressed related to current practice:

3. What logical implementations and software packages are currently used by practitioners?

4. Do the implementations fail to detect the errors or do the implementations introduce additional errors?

Finally, in order to address the issue covered by the previous research:

5. How might these techniques and analyses be applied to the NTSB database text fields for an exploratory analysis of factors contributing to accident fatality?

RESEARCH APPROACH.

The research consists of five activities, executed in sequence:

1. Literature survey for background information and reference.

2. Identification of available software and techniques—collection of data on existing text mining and text analysis tools.

3. Software platform preparation and evaluation—acquisition and installation of available tools and software packages, evaluation of appropriateness to research task.

4. Data parsing and preliminary analysis—data preparation, text mining and information extraction.

5. Data analysis and reporting—analysis of the data and documentation of the process and results in this report.

LITERATURE OVERVIEW

One of the objectives of this report study is to research the literature related to the use of text analysis techniques and their software implementations. The included literature focuses on providing a general overview of issues related to the use of text analysis. Literature was categorized from three perspectives:

1. A general research perspective that includes broad overviews of techniques in text analysis.

2. A study of the application of text analysis in nonaviation industries.

3. A focus on the application of text mining in safety or transportation areas.

The articles included in the annotated bibliography address these perspectives and serve to further inform the research process. In addition to the academic publications, “Mining Aviation Safety Reports Using Predictive Word Sequences” (Melby, 2011) was invaluable in explaining the algorithms behind the ASIAS Information Retrieval and Extraction System (AIRES) software. This report was greatly assisted by Dr. Melby’s article, as the construction of the AIRES software provided an initial base from which the exploration into other techniques was launched. The perfection of text analysis techniques has long been a goal in the computer and information science fields. This can be primarily attributed to the abundance of text data available because of the organizational and psychological preference for written reports. A foundational article within the field of text analytics (Salton et al., 1975) demonstrates that words and patterns can be ranked in terms of how well they are able to discriminate between documents of a collection. This discrimination value analysis is computationally simple and allows for the development of analytical techniques based on the initial words determined by the process. Similar research was conducted by Delen & Crossland (Delen and Crossland, 2008). The authors established the importance of text analysis using a case study and developed a methodology they termed “IDEF” for the purposes of exploring the text data present in the case. Their research demonstrated the usefulness of text analysis, even in circumstances where structured data were available to represent a subset of the total information concerning the analysis target. Acknowledging that text analysis is useful, applicable to the research approach, and computationally manageable, several techniques were reviewed to determine their suitability for the purposes of this report. Four of the commonly used methodologies in the extraction of quantifiable data from text banks were identified (Lee et al., 2010). These methodologies were compared and evaluated by the authors to establish their effectiveness and applicability to types of problems. This comparison was used to inform the selection of an inverse document frequency term list as the mining tool in the STATISTICA analysis discussed below. With the mining technique decided, the next stage of the research approach called for a review of available software with which to accomplish the mining and analysis of data. The software packages identified (Zhang and Segall, 2010) were evaluated by the research team and a comparison of their features was compiled. During this process, an additional software package not identified by Zhang and Segall, STATISTICA, was discovered to meet the necessary criteria for use in the analysis.

Prior to conducting the analysis, an investigation was made into previous applications of text mining and analysis in aviation and transportation safety applications. Among the articles identified by this process were “Applied Hermeneutics and Qualitative Safety Data” (Wallace et al., 2003) and “Mining and Tracking Massive Text Data” (Jeske and Liu, 2007). Both detailed efforts to use text mining in the analysis of transportation safety data. The Wallace et al. article was approached from a traditional perspective, using a modified quantitative-qualitative system. This approach was chosen largely because the data contained within the available database was confidential and, thus, positivist measures could not be used. Despite this limitation, the article provided solid information regarding the nomenclature of the text mining field and text categorization and interpretation. The Jeske and Liu article was perhaps the one most closely related to the fundamental purpose of this report, as it details an analysis of FAA aviation safety reports. Jeske and Liu’s use of a naive Bayesian classifier is similar to this report’s use of K-Means Clustering and helped to inform the choice of which text fields of the database to use in the analysis.

When conducting the initial literature survey, it was found that several papers related to the application of intelligent algorithms, such as genetic algorithms and neural networks to the problem of text mining and classification. These approaches have a separate set of potential benefits and drawbacks and were beyond the scope of this report. However, two articles in particular (Kloptchenko et al., 2004 and Tseng et al., 2005) demonstrated that the application of Self-Organizing Maps (SOM) to text analysis problems can yield meaningful results. Kloptchenko et al. identify the potential for combining structured and unstructured items in a mosaic approach and demonstrate that the SOMs can categorize text data for this purpose. Using the neural networks to cluster like terms for analysis, Tseng et al. further demonstrate text mining in a transportation safety capacity. After the literature survey had been completed, the research methodology was developed for the following purposes: to explore the text data for correlations between text patterns and accident fatality, to provide a basis for comparison to the AIRES software provided by the MITRE Corporation, and to verify the results of previous analyses conducted by the authors using logistic regression.

METHODOLOGY

THE AIRES.

The negative consequences of aircraft accidents on manufacturers, operators, the industry as a whole, and the general public have been well established. To promote the open exchange of safety information to facilitate continuous improvement in aviation safety, the FAA has developed the ASIAS system. This system enables users to perform integrated queries across various aviation safety databases.

Models and insights developed using this system are then used throughout the industry to generate improvements in safety practices. In collaboration with the ASIAS initiative, the MITRE Corporation has begun work on a software program to solve the data sufficiency problems that exist in the structured fields of the ASIAS databases. The AIRES addresses the issue of insufficient information within the structured fields by conducting an analysis of the more complete narrative report fields present within the databases. At a high level, the software functions by comparing positively and negatively labeled records to discover words and word sequences that have predictive power over a desired dependent variable.

The proportion of contributing factors to overall incident types illustrates one of the strengths of the AIRES algorithm over frequency-based methods. The strength of the AIRES approach is its method for addressing word combinations where gaps exist between the words in a phrase. For example, in the phrase "The operator proceeded to lose control of the aircraft before a collision with terrain," one could argue that the relevant words are lose, control, collision, and terrain. However, as they are separated by independent, uncorrelated words, a traditional analysis would only include them individually. The AIRES approach is capable of ignoring the words between each relevant word and constructing the phrase "lose control collision terrain," which may have better predictive power than any of those words individually.

The results display created by the AIRES tool is grouped into eight columns in total, namely: Pattern, Information Gain, Precision, Recall, Weighted F-Measure, Lift, Positive Reports, and Total Reports. The pattern column contains the relevant word or group of words corresponding to the numeric results in the other column. The lift and report measures are useful in determining how many reports contain the word or group of words. A highly predictive word sequence that appears in a minority of reports is of limited usefulness for determining trends in a broad sample of accidents. Information gain, precision, recall, and weighted F-measure are used in determining the significance of the pattern. Information gain represents the increase in the accuracy of a predictive model that includes the selected term. Precision is a measure of the number of retrieved instances that are relevant. Recall is the number of relevant instances that are retrieved. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer instances of the pattern have been missed. The F-measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting. Practitioners vary in their opinions of the relative usefulness of precision and recall by application, resulting in different Weighted F-Measures.

STATISTICA.

MITRE's AIRES package is an automated high-level implementation of its underlying algorithm. A user with little technical knowledge can be quickly trained in the use of the software and begin applying their domain knowledge almost immediately. By contrast, the other software packages reviewed for use in this report are considerably more manual- and knowledge-intensive. Based on a review of the most commonly used text mining software, as identified in the Zhang and Segall paper, the research team reached the decision to conduct the analysis in the STATISTICA software package. To evaluate the robustness of both the model generated for the previous report and the AIRES algorithm, the STATISTICA package was used to perform an exploratory analysis. This process involves several steps, beginning with reducing the problem into more manageable terms through the use of frequent words. This reduced word set was then put through a technique called "singular value decomposition" to allow for the examination of trends within the documents. This process generates many concepts for a given document, with each concept representing an amount of variation within the text.

Processing each additional concept consumes computational resources. As such, the optimal case is to use the minimum number of concepts that capture a great amount of the variance. To decide what number of concepts meets these criteria, a scree plot is generated, which illustrates the percentage of variance explained by each additional concept. The "elbow" of this plot is the point at which the increase in variance explained levels off; therefore, this point can be used to determine the correct number of concepts for inclusion. The bulk of the informative variation (non-scattered data) is captured by the first three concepts. These initial components have been selected for use in the analysis. The components selected by this process can be used to generate word coefficients for use in Principal Components Analysis (PCA). The PCA demonstrates which terms within a document represent the greatest variability with the corpus as a whole. There are multiple methods for arriving at a plot of this variability, two of which are included in this report.

The PCA technique is useful in identifying trends for subsequent analysis and review by field experts, due in part to the fact that it is a graphical approach requiring little in the way of statistical background. However, the graphical nature of the analysis forces the human interpreter to identify the significance of the clusters, unlike the AIRES method. The use of a K-Means algorithm is similar to the AIRES method in that it automatically identifies clusters of related items. Unlike the AIRES approach, it clusters the data based on similar documents from which important terms are then identified. To identify these terms, documents that are representative of their respective clusters are sampled based on proximity to the cluster mean.

RESULTS

NATIONAL.

This section relates to results derived from the sum of all documents within the selected portion of the NTSB aircraft accident database. For each entry in the database, the cause narrative report has been used to determine words and phrases that may be indicative of accident fatality. In general, results on a national level were too scattered to yield actionable results and, therefore, the analysis was broken into regional subsections. It is likely that the results on a national level were populated by a larger number of smaller patterns, as certain effects occur exclusively or more frequently in different environments.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0 with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting. The weighting of this measure in table 1 demonstrates the diversity of the NTSB database in that no individual term has a very high significance as measured by the F-Measure. The figure represents the analysis of high-fatality patterns within all national accident cause text reports. The patterns correspond to the top 20 words or phrases associated with fatal aircraft accidents and the results that are produced demonstrate several useful patterns. Although the patterns specified in table 1 are of limited usefulness individually, together they form a description of the fatality contributing factors. For instance, the patterns—instrument, meteorological, IMC, Visual Flight Rules (VFR), continued flight, and into—suggest that flights into adverse weather possibly under VFR rules are highly correlated to accident fatality.

Table 1. The AIRES Results, National

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
instrument	0.034785339	0.828322017	0.076612541	0.09360135
into	0.034527965	0.663175303	0.107921414	0.129628033
meteorological	0.022699259	0.821325648	0.051134834	0.062938916
low	0.019843652	0.538133333	0.090517628	0.108581021
spin	0.019439383	0.801574803	0.04566251	0.056276673
vfr	0.01846783	0.796416938	0.043868305	0.054090527
weather	0.017509668	0.528906697	0.082892258	0.099708644
spatial	0.017372428	0.887096774	0.034538441	0.042756874
disorientation	0.016171269	0.828865979	0.036063515	0.044594325
mountainous	0.014058431	0.725752508	0.038934242	0.048023724
maneuvering	0.013674969	0.593373494	0.053018749	0.064825377
adverse	0.012279979	0.636118598	0.042343231	0.052062652
pilots, flight	0.012135432	0.669796557	0.038395981	0.047316868
night	0.011852034	0.519398258	0.058849915	0.071536062
imc	0.011674422	0.783783784	0.028617565	0.035448383
continued, flight	0.011153734	0.739514349	0.030052929	0.03718834
impairment	0.01058536	0.798295455	0.025208576	0.031263907
continued	0.010323867	0.66427289	0.033192787	0.040979067
maintain, altitude	0.009208638	0.589259797	0.036422356	0.044835126
dark	0.008347821	0.529481132	0.040279896	0.049410159

The AIRES tool demonstrates that the top 20 patterns do not account for much of the variability within the reports. This indicates that the accidents are best described by a large number of different patterns, rather than a small number of similar patterns. The patterns with the highest F-Measure include: into, low, weather, instrument, and night. While some preliminary conclusions could be drawn based on these terms, their low level of information gain would mean that the predictive power of a model based on this information would be low. The usefulness of this information devoid of context is also questionable, as the top results mean very little to observers on their own; that the results have this quality necessitates further analysis.

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, a PCA is conducted on the data to identify word clusters that cause variance within the corpus of national documents. The higher a word is on either axis, the more variation it causes within the document. Words that are grouped together are related as illustrated by the red circles within the figures. The term "concept" in this case refers to the amount of total variation included in the data. For example, if an analysis requires 30 concepts to explain 100% of the variation, then concepts 1 and 2 are the largest two single contributors to the variation. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Notably absent from these results is the top return of the AIRES analysis "into." This is due to the STATISTICA package's implementation of word filtering based on common English words. Further analyses of this type are available in the appendices.

Figure 1 shows words that are responsible for the majority of variability in aircraft accident reports that were prepared for the national level. Concepts 1 and 2 are the largest single contributors to the variability. Where the words appear relative to these axes demonstrates the importance of these words to the concepts and the document as a whole (i.e., the higher a word is on either axis, the more variation it causes within the document). The words within each of the red circles represent the words that are grouped together.

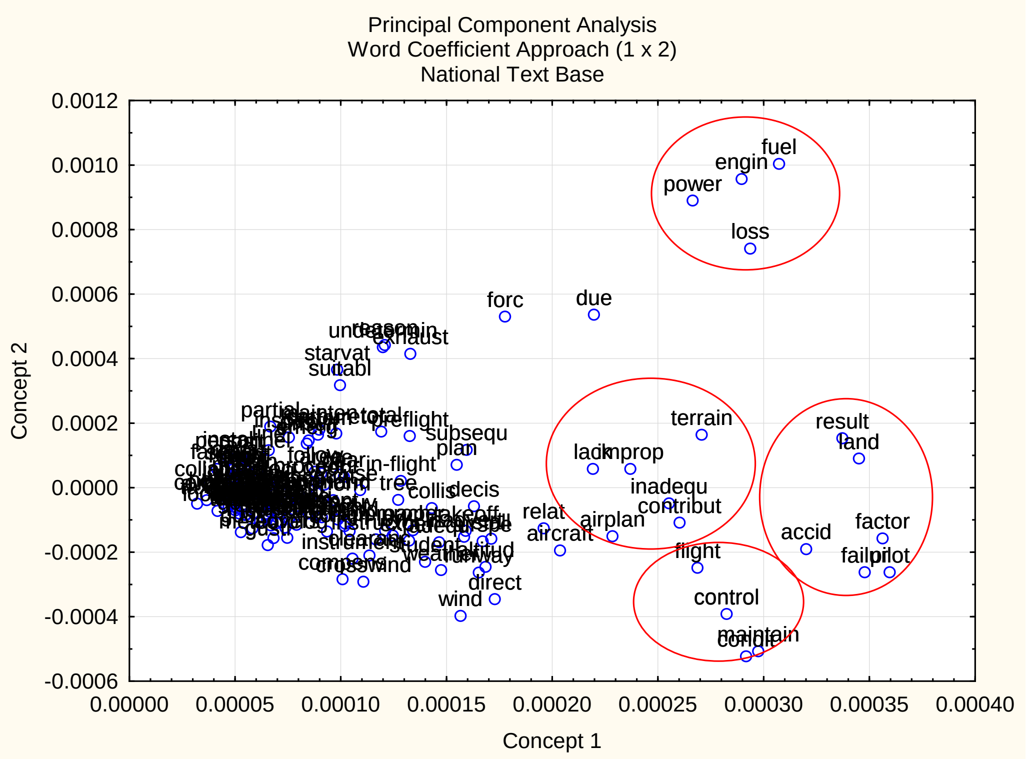

Figure 1. The PCA, Coefficients, National

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents; additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The use of a K-Means algorithm is similar to the AIRES method in that it automatically identifies clusters of related items. Unlike the AIRES approach, it clusters the data based on similar documents from which important terms are then identified. To identify these terms, documents from the clusters are chosen based on how well they represent their cluster.

The representative documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide the context the pattern follows. On a national level, there were three clusters of significance composed of several subcomponents each. A cross-tabulation of these clusters with accident fatality condition as well as example text from each cluster are available in the appendices.

The first fatality indicating cluster concerns inadequate weather evaluation leading to visual flight into IMC, particularly darkness and fog/low cloud ceilings. The second such cluster indicates that a loss of engine power resulting in a failure to maintain airspeed is responsible for many fatal accidents on a national scale. The third cluster indicates that failure to maintain airspeed resulting in a stall separate from any engine or power loss is also a leading cause of fatal accidents. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 1: WESTERN PACIFIC.

This section relates to the Western Pacific region, which consists of Arizona, California, Hawaii, and Nevada. Factors leading to fatal accidents within this region were defined by the analyses and were closely related to the national-level results.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, in table 2, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to stalling, airspeed, and altitude, as well as such adverse meteorological conditions as darkness, night, and weather. Again the words “into” and “low” are included as relevant patterns, which are unfortunately lacking in context. The Weighted F-Measures for the region are considerably higher than on a national level, indicating that the region has causes that are partially unique among its peers.

Table 2. The AIRES Results, Region Western Pacific

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
into	0.046517307	0.724489796	0.133899104	0.159981974
stall	0.022312781	0.525562372	0.121169260	0.143207400
instrument	0.047017062	0.877118644	0.097595474	0.118692661
weather	0.028031781	0.671140940	0.094295144	0.113869278
low	0.020676991	0.587692308	0.090051862	0.108411852
mountainous	0.031591610	0.774336283	0.082508251	0.100459242
meteorological	0.033605826	0.868571429	0.071664309	0.087769950
maneuvering	0.015659857	0.614678899	0.063177746	0.076993795
vfr	0.028524757	0.866666667	0.061291843	0.075283762
spin	0.029724378	0.889655172	0.060820368	0.074747943
pilots, flight	0.019634383	0.778571429	0.051390853	0.063195733
night	0.011259521	0.579787234	0.051390853	0.062845941
with, terrain	0.010920473	0.579234973	0.049976426	0.061151494
adverse	0.017328299	0.785123967	0.044790193	0.055200465
dark	0.010467029	0.609271523	0.043375766	0.053271569
maintain, altitude	0.013287297	0.716666667	0.040546912	0.049976755
terrain, clearance	0.014085615	0.750000000	0.039603960	0.048859935
airspeed, while	0.011417861	0.696428571	0.036775106	0.045369940
spatial	0.018227569	0.905882353	0.036303630	0.044929397

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. Words that are higher on either axis represent more variation within the document, and words that are collocated are related. The PCA identified several clusters of words, many of which were not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and are, therefore, useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

The PCA in figure 2 shows the importance of the same terms in the AIRES algorithm, as well as pilot control and the loss thereof. Power, engine, and loss are also identified, which is consistent with the results on the national level. Many words that, at first glance, would not seem to be relevant to fatality prediction are shown to generate significant document variability. Among these are pilot, aircraft, and airplane. This may indicate that accidents in which the pilot was at fault were more often fatal than those caused by other factors. The relevance of the terms "aircraft" and "airplane" may likewise indicate that equipment failures also contribute.

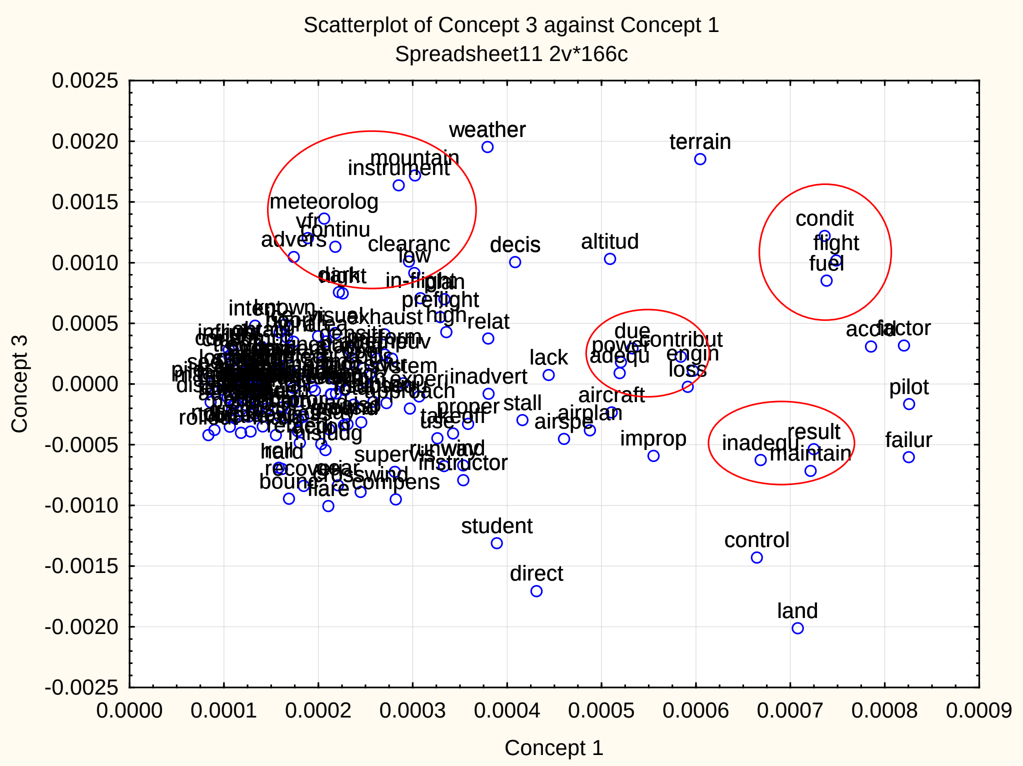

Figure 2. The PCA Coefficients (1 x 3), Region Western Pacific

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents; additional techniques can be used to identify the most descriptive clusters. The use of a K-Means algorithm is similar to the AIRES method in that it automatically identifies clusters of related items. Unlike the AIRES approach, it clusters the data based on similar documents from which important terms are then identified. To identify these terms, documents that are representative of their respective clusters are sampled based on proximity to the cluster mean. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

As with the national-level analysis, the K-Means clustering process identifies three document clusters that are primarily indicative of fatal accidents. The first cluster included pilot failure to maintain adequate clearance altitude in mountainous or hilly terrain. The second cluster involved pilot failure to maintain airspeed resulting in a stall and subsequent uncontrolled collision with terrain. A minority of cases also involved the use of both over-the-counter and illicit drugs. The third cluster underscored the dangers of visual flight into IMC. The greater granularity of the regional report allows some insight into the causes of flight into these conditions and their eventual fatality, including: controlled flight into rising terrain, insufficient instrument training, pressure to adhere to a particular flight route, failure to obtain a weather briefing, and impairment by controlled substances. This is consistent with the previous report's results for the region that included VFR flight into IMC as the leading single factor, followed by airspeed as the second leading factor. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 2: SOUTHWEST.

This section relates to the Southwest region, which consists of Arkansas, Louisiana, New Mexico, Oklahoma, and Texas. Results within the region were generally characterized by stalling, adverse weather conditions, and inadequate terrain clearance.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, whereas a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 3, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to stalling, airspeed, and altitude, as well as adverse meteorological conditions—including the effects of these conditions, such as spatial disorientation and impairment. The Weighted F-Measure indicates a less robust confidence in the analysis than other regions, but still greater than that of the sum of all documents nationally.

Table 3. The AIRES Results, Region Southwest

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
low	0.026412234	0.600000000	0.099861304	0.119840213
into	0.025859083	0.657608696	0.083911234	0.101646505
weather	0.016205132	0.517543860	0.081830791	0.098398933
instrument	0.029059521	0.830357143	0.064493759	0.079081633
adverse	0.012722216	0.631067961	0.045076283	0.055356839
spin	0.017482522	0.800000000	0.041608877	0.051343488
maneuvering	0.007823227	0.522935780	0.039528433	0.048494130
disorientation	0.015206303	0.767123288	0.038834951	0.047936997
failure, airspeed	0.007696035	0.523364486	0.038834951	0.047659574
spatial	0.019226007	0.900000000	0.037447989	0.046328071
while, maneuvering	0.008049521	0.545454545	0.037447989	0.046020112
pilots, flight	0.009317190	0.595505618	0.036754508	0.045245006
meteorological	0.016826999	0.852459016	0.036061026	0.044604563
at, altitude	0.008425483	0.602564103	0.032593620	0.040198426
vfr	0.013554081	0.807017544	0.031900139	0.039484979
impairment	0.011115531	0.725806452	0.031206657	0.038593482
stall, factors	0.008177629	0.623188406	0.029819695	0.036833990
fog	0.009141608	0.683333333	0.028432732	0.035175017
sufficient	0.006026690	0.557142857	0.027045770	0.033401850

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and are thus useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

The principal components analysis in figure 3 illustrates the importance of the same terms as well as the need for the pilot to maintain control in these unfavorable conditions. Loss of power and pilot failure to comply with procedure are closely related within this region of power loss and terrain collision. Likewise, mentions of lighting condition and weather are somewhat related to identification of instrument conditions during the flight.

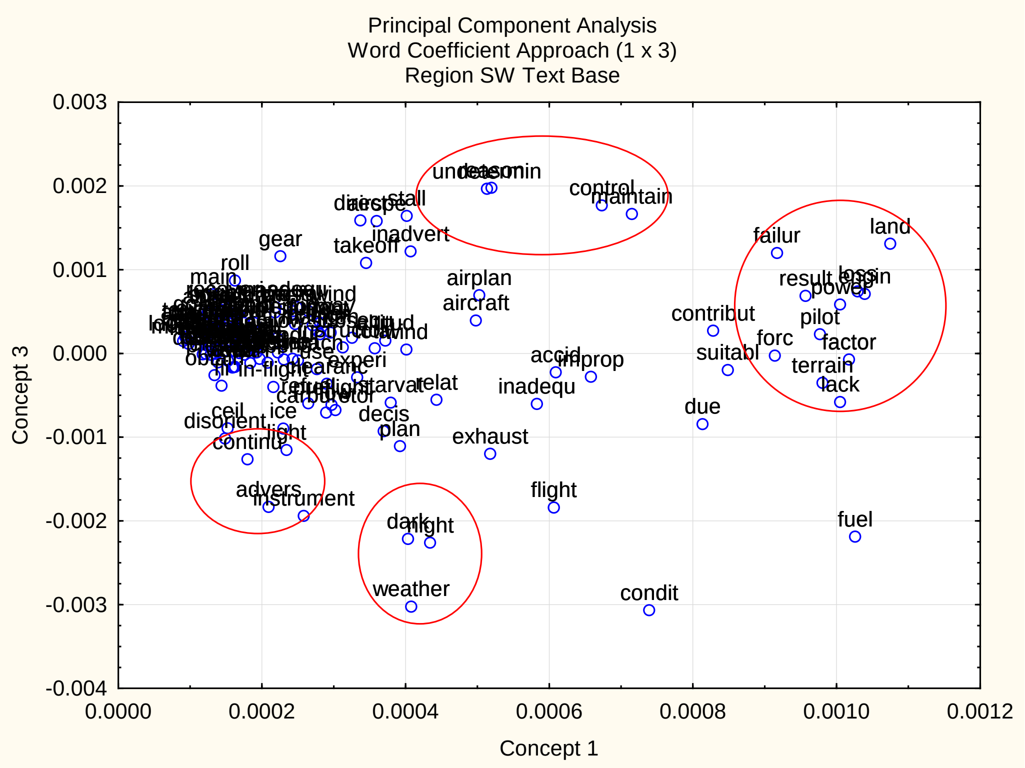

Figure 3. The PCA Coefficients (1 x 3), Region Southwest

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents. Additional techniques can also be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

The region features three prominent clusters that are typically correlated to fatality. The first cluster includes intentional flight into instrument conditions and subsequent spatial disorientation, particularly where the adverse conditions include darkness and thunderstorms.

The second cluster identifies failure to maintain adequate terrain clearance at night as being highly likely to cause a fatal accident. The third cluster involves failure to maintain adequate airspeed resulting in a stall. Some of the language within the prototype text is highly similar and formulaic, possibly indicating a single author for many of the documents or a standard format for the report. The previous report identified aircraft control, airspeed, VFR flight into IMC, and insufficient clearance as the leading contributors to accident fatality within the region.

REGION 3: ALASKAN.

This section relates to the Alaskan region, which consists of Alaska. The region as a whole does not contain well-characterized accident forms, as is indicated by the information gain of the most predictive term patterns within the regional corpus. This is perhaps due to the variety of challenging flight conditions present in the state of Alaska and the relatively low number of reports, given that the region contains only one state. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 4, highlight several terms and phrases, which are highly correlated to fatal accidents. Among these are several terms relating to stalling, airspeed and altitude, and adverse meteorological conditions (particularly low cloud ceilings). The Weighted F-Measure is low for these results, bordering on that of the consolidated national document corpus. The information gain from the top patterns is also low, indicating that fatal accidents within the region are not well characterized by a small number of patterns.

Table 4. The AIRES Results, Region Alaskan

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
mountainous, terrain	0.034126374	0.596153846	0.087570621	0.105585831
into, conditions	0.027848149	0.527272727	0.081920904	0.098572400
vfr	0.030715794	0.595744681	0.079096045	0.095693780
instrument	0.030548113	0.613636364	0.076271186	0.092465753
meteorological	0.026433611	0.600000000	0.067796610	0.082417582
resulted, inadvertent, stall	0.016857227	0.548387097	0.048022599	0.058742225
Both	0.015364311	0.533333333	0.045197740	0.055325035
maintain, airspeed, while	0.017924726	0.652173913	0.042372881	0.052119527
imc	0.013562757	0.538461538	0.039548023	0.048543689
ceilings	0.015453136	0.650000000	0.036723164	0.045264624
adequate, while	0.011771839	0.545454545	0.033898305	0.041724618
Clouds	0.017396808	0.846153846	0.031073446	0.038488453
each	0.017396808	0.846153846	0.031073446	0.038488453
maintain, while, maneuvering	0.010880970	0.550000000	0.031073446	0.038300836
at, altitude	0.010299521	0.523809524	0.031073446	0.038274182
accident, low	0.009993703	0.555555556	0.028248588	0.034867503
airspeed, maneuvering	0.009993703	0.555555556	0.028248588	0.034867503
Inadequate, visual	0.009404901	0.526315789	0.028248588	0.034843206
at, low	0.009404901	0.526315789	0.028248588	0.034843206

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and are, therefore, useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

In figure 4, the PCA is not particularly conclusive beyond the AIRES algorithm other than to suggest that pilot control and the takeoff phase of flight are important indicators. Engine power loss as the result of fuel starvation continues to be identified as a meaningful cluster; however, it is not present in a significant enough portion of the documents to yield acceptable predictive accuracy.

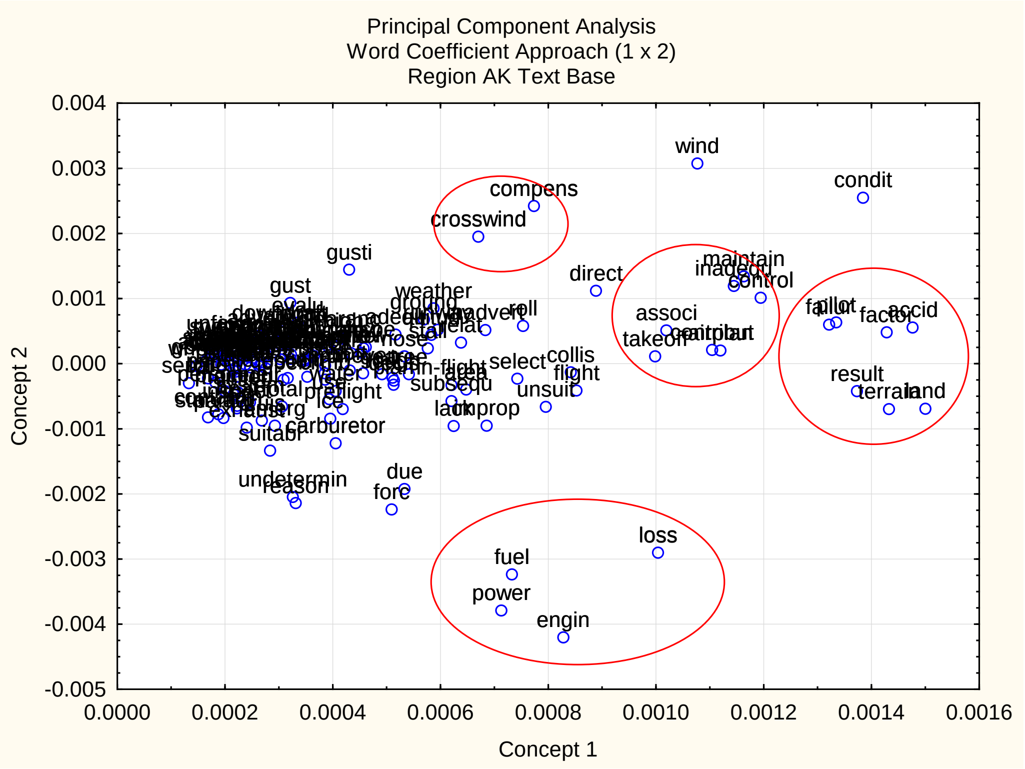

Figure 4. The PCA Coefficients (1 x 2), Region Alaskan

K-MEANS CLUSTERING. The data contain words and clusters of words, which can be used to classify the documents. Additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but also provide context.

The prototype text for the region is separated into two main clusters as per the analysis conditions. Although 22 total clusters were identified, more than any other region in this report, most of these clusters were not highly correlated to accident fatality. The two clusters that did have a high correlation contained information as follows. The first cluster emphasizes the potential fatality of visual flight into IMC, particularly low cloud ceilings and fog. The second cluster identifies failure to maintain airspeed, resulting in a stall at low altitude as a highly fatal series of flight factors. The cluster concerning visual flight into IMC contains a majority of fatal accidents; however, the low airspeed stall cluster contains only a relative majority and therefore may not have represented the same level of predictive accuracy. The previous report identified VFR flight into IMC, airspeed, in-flight planning/decision, and stalling as the leading contributors to accident fatality within the region. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 4: CENTRAL.

This section relates to the Central region, which consists of Iowa, Kansas, Missouri, and Nebraska. The fatal accidents with the Central region are generally characterized by stalling and visual flight into IMC. Additionally, the region is subject to a preponderance of loss of control at night while under the influence of mind-altering substances.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 5, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to airspeed, IMC, stalling, and pilot maneuvers. The Weighted F-Measure for these results is high relative to the national level, indicating a highly clustered document corpus. However, the information gain from each individual pattern is relatively low indicating that the corpus is also diverse.

Table 5. The AIRES Results, Region Central

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
low	0.030015525	0.513043478	0.125531915	0.147869674
into	0.046483387	0.690476190	0.123404255	0.147657841
weather	0.027231819	0.537634409	0.106382979	0.126710593
instrument	0.045472585	0.843137255	0.091489362	0.111341274
maintain, airspeed	0.017424480	0.523076923	0.072340426	0.087403599
known	0.018496901	0.588235294	0.063829787	0.077679959
inadvertent, stall	0.014021643	0.529411765	0.057446809	0.069911963
meteorological	0.025674297	0.833333333	0.053191489	0.065445026
spin	0.021929065	0.793103448	0.048936170	0.060240964
disorientation	0.024503347	0.880000000	0.046808511	0.057742782
failure, maintain, adequate	0.010692898	0.525000000	0.044680851	0.054687500
flying	0.014772458	0.666666667	0.042553191	0.052356021
pilots, flight	0.014070326	0.645161290	0.042553191	0.052328624
fog	0.014393345	0.678571429	0.040425532	0.049790356
adequate, airspeed	0.010205895	0.542857143	0.040425532	0.049608355
maneuver	0.012170977	0.653846154	0.036170213	0.044596013
clearance, terrain	0.013110172	0.750000000	0.031914894	0.039473684
maneuvering	0.009412207	0.600000000	0.031914894	0.039370079
pilots, failure, adequate	0.008852903	0.576923077	0.031914894	0.039349423

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and are, therefore, useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

In figure 5, the principal components of the documents for the central region indicate that failure to maintain control of the aircraft is a leading element within the region. Engine power loss due to fuel starvation continues to be indicated as responsible for a minority of fatal accidents.

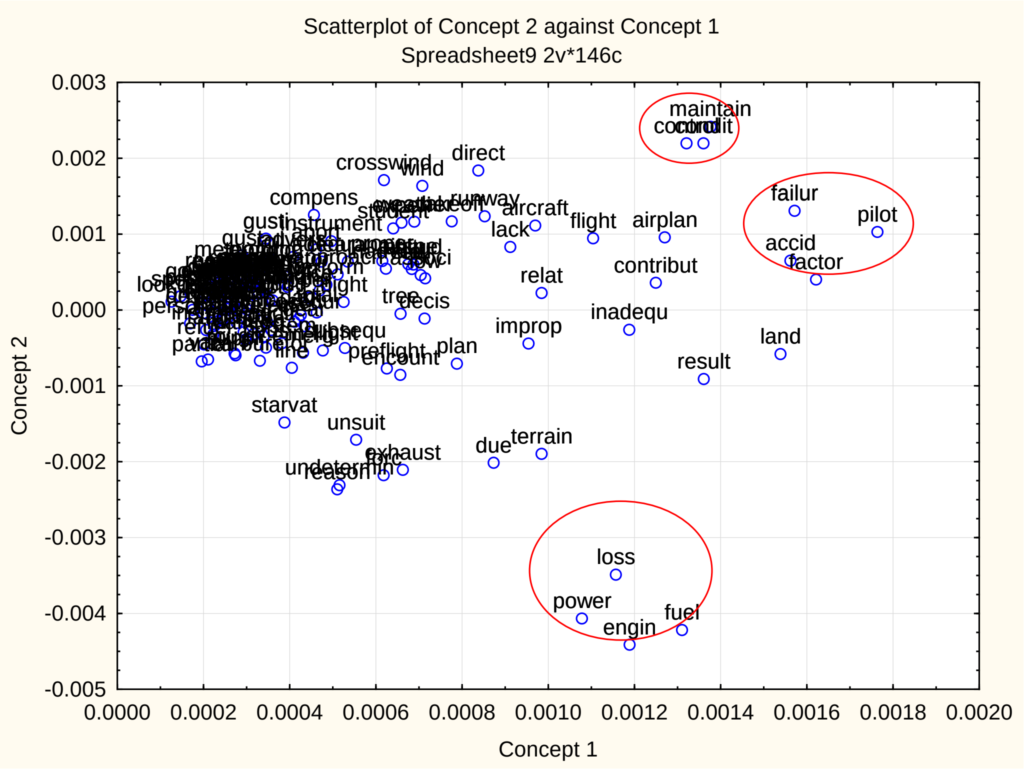

Figure 5. The PCA Coefficients (1 x 2), Region Central

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents. Additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, they also provide context.

The prototype text for the region is separated into three main clusters as per the analysis conditions. The first cluster indicates the failure to maintain altitude in areas with a high-terrain clearance and executing low-altitude maneuvers within these conditions is often fatal. The second cluster corresponds to accidents involving failure to maintain control of the aircraft during flight in adverse weather. These conditions can be exacerbated by the use of alcohol and marijuana. The third cluster concerns failure to maintain airspeed, resulting in a stall. The previous report identified aircraft control, airspeed, VFR flight into IMC, and poor preflight planning as the leading contributors to accident fatality within the region. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 5: GREAT LAKES.

This section relates to the Great Lakes region, which consists of Illinois, Indiana, Michigan, Minnesota, North Dakota, Ohio, South Dakota, and Wisconsin. The region's fatal accidents are generally characterized by failure to maintain airspeed, failure to maintain altitude, and visual flight into IMC.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 6, highlight several terms and phrases that are highly correlated to fatal accidents. These are terms relating to stalling, airspeed, and altitude, as well as adverse meteorological conditions, including the effects of these conditions, such as spatial disorientation and impairment. General English terms, including “continued,” “into,” and “low” are contained within the selected patterns, reducing the usefulness of the results in accident prediction. The Weighted F-Measures for the region were typical of the national results with the information gain being relatively low, indicating a diverse group of patterns within the document corpus.

Table 6. The AIRES Results, Region Great Lakes

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
into	0.028157119	0.615763547	0.091575092	0.110365531
instrument	0.031187904	0.814159292	0.067399267	0.082540822
spin	0.022670134	0.762886598	0.054212454	0.066582689
airspeed, which	0.011041409	0.539130435	0.045421245	0.055605381
adverse	0.011058535	0.550458716	0.043956044	0.053869635
adverse, weather	0.012426510	0.604166667	0.042490842	0.052195824
maneuver	0.010938414	0.564356436	0.041758242	0.051249775
meteorological, conditions	0.018333792	0.800000000	0.041025641	0.050632911
meteorological	0.017577776	0.777777778	0.041025641	0.050614606
maneuvering	0.012729157	0.642857143	0.039560440	0.048701299
disorientation	0.014414704	0.735294118	0.036630037	0.045224313
maintain, altitude	0.008827467	0.537634409	0.036630037	0.045020710
spatial	0.016521063	0.816666667	0.035897436	0.044384058
weather, conditions	0.008050024	0.515789474	0.035897436	0.044104410
airspeed, resulted	0.008039141	0.528089888	0.034432234	0.042349973
failure, airspeed	0.007284890	0.505494505	0.033699634	0.041433976
altitude, clearance	0.007256389	0.524390244	0.031501832	0.038794659
pilots, flight	0.008217384	0.600000000	0.028571429	0.035294118
continued	0.006793130	0.542857143	0.027838828	0.034358047

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, a PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and are, therefore, useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

In figure 6, the PCA illustrates the importance of the same terms as well as power loss as a result of engine failure or loss of fuel. Pilot failure to maintain control also represents variability within the accidents occurring in the Great Lakes Region.

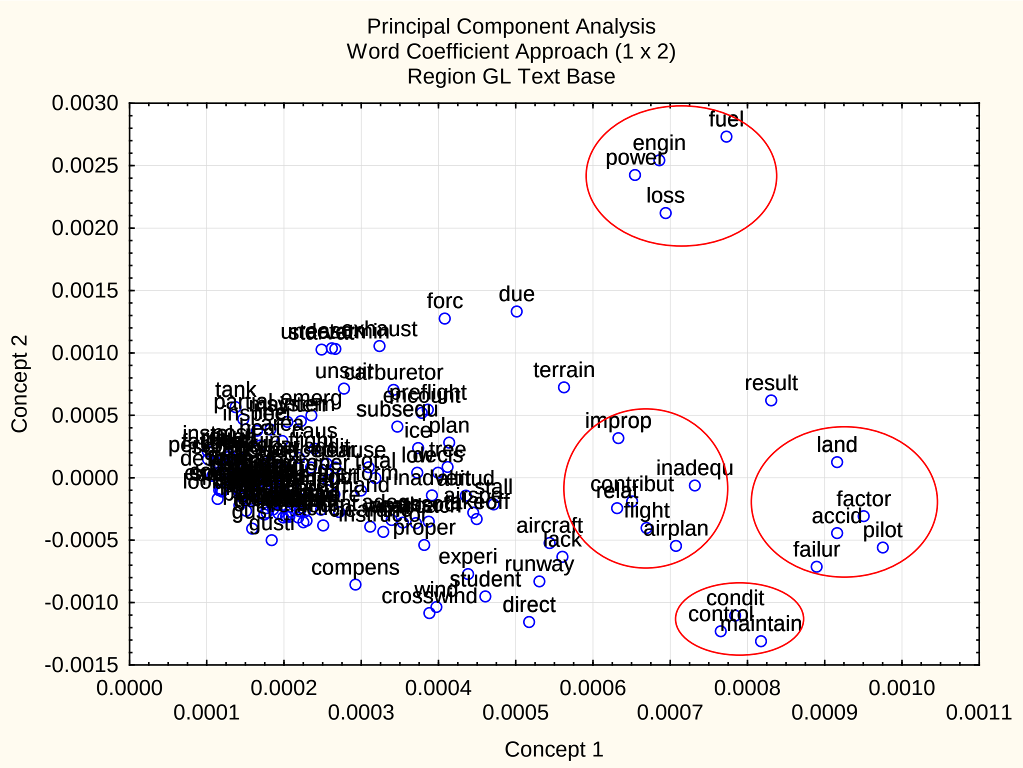

Figure 6. The PCA Coefficients (1 x 2), Region Great Lakes

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents. Additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

The prototype text for the region is separated into three main clusters as per the analysis conditions. The first cluster indicates that failure to maintain adequate airspeed resulting in a stall was a significant contributor to accident fatality. The second cluster was poorly characterized and included dark night conditions, failure to maintain proper altitude, and alcohol-related pilot impairment. The third cluster highlighted the importance of avoiding visual flight into IMC and the hazards of pilot overconfidence in such conditions. Lacking instrument certification and instrument meteorological condition experience were identified as lesser factors. The previous report identified aircraft control, airspeed, VFR flight into IMC, and insufficient clearance as the leading contributors to accident fatality within the region. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicate that the different approaches achieve a similar result, though they are not directly comparable.

REGION 6: NORTHWEST MOUNTAIN.

This section relates to the Southwest region, which consists of Colorado, Idaho, Montana, Oregon, Utah, Washington, and Wyoming. Accidents within the region are primarily characterized by stalling, adverse weather conditions, and failure to maintain adequate terrain clearance.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 7, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to spinning, airspeed, and mountainous terrain clearance, as well as adverse meteorological conditions. The standard English words “into,” “low,” and “from” are included, and are of little inherent predictive value. VFR occurs twice within the results—once as a unique pattern and once as part of a pattern with flight.

Table 7. The AIRES Results, Region Northwest Mountain

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
into	0.038897629	0.646048110	0.124503311	0.148475754
low	0.026255518	0.540372671	0.115231788	0.136749450
mountainous	0.030808815	0.676328502	0.092715232	0.112053786
maneuvering	0.020430069	0.594871795	0.076821192	0.093023256
instrument	0.027735464	0.806722689	0.063576159	0.077934730
night	0.016337255	0.585365854	0.063576159	0.077369439
weather, conditions	0.012404461	0.510869565	0.062251656	0.075514139
terrain, factors	0.012688125	0.522727273	0.060927152	0.074002574
maintain, adequate, airspeed	0.010712947	0.509316770	0.054304636	0.066118368
meteorological	0.023101533	0.801980198	0.053642384	0.065950171
adverse	0.016687405	0.661157025	0.052980132	0.064924525
adverse, weather	0.017195430	0.700934579	0.049668874	0.061005368
clearance, terrain	0.014729044	0.643478261	0.049006623	0.060113729
dark	0.012762446	0.592000000	0.049006623	0.060016221
spin	0.022370696	0.839080460	0.048344371	0.059572385
vfr	0.016507590	0.695238095	0.048344371	0.059397884
from, terrain	0.012427201	0.612612613	0.045033113	0.055275565
vfr, flight	0.017475252	0.767441086	0.043708609	0.053868756
clearance, from, terrain	0.013230552	0.647058824	0.043708609	0.053728427

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and, thus, are useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents. The PCA, in figure 7, illustrates the importance of the same terms, as well as power loss due to engine failure or fuel condition.

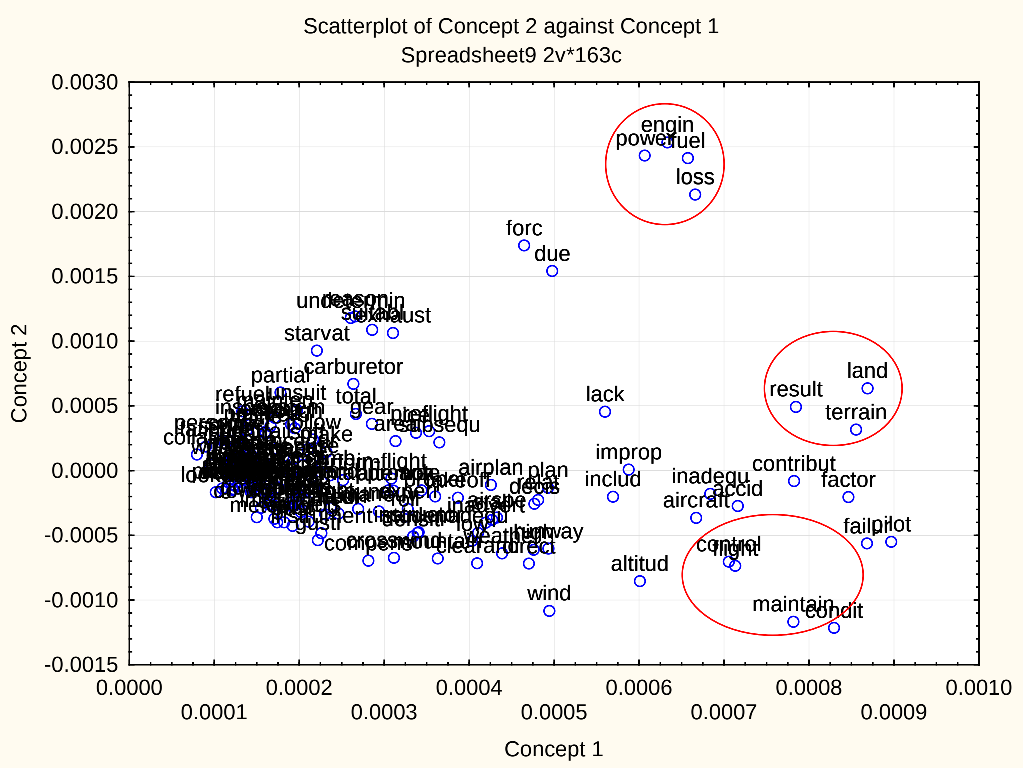

Figure 7. The PCA Coefficients (1 x 2), Region Northwest Mountain

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents. Additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

The prototype text for the region is separated into three main clusters as per the analysis conditions. The first cluster identifies visual flight into IMC, particularly in cases where these conditions involve low cloud ceilings, rain, and fog. The second cluster concerns failure to maintain proper altitude either in cases of rising terrain or on approach to the landing airport. The third cluster involves stalling due to a failure to maintain proper airspeed. This cluster is cross-correlated with the previous cluster in that some accidents within it were also at an improper altitude, increasing the severity of their stalling conditions. The previous report identified in-flight planning and decisions, VFR flight into IMC, aircraft control, and airspeed as leading issues. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 7: NEW ENGLAND.

This section relates to the New England region, which consists of Connecticut, Massachusetts, Maine, New Hampshire, Rhode Island, and Vermont. Fatal accidents within the New England region typically involve both stalling or marginal weather conditions, and subsequent contributory factors.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 8, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to stalling and altitude, as well as adverse meteorological conditions and loss of aircraft control.

Table 8. The AIRES Results, Region New England

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
altitude	0.025295508	0.510000000	0.124694377	0.146889401
low	0.029515489	0.650000000	0.095354523	0.114976415
into	0.023295086	0.575757576	0.092909535	0.111633373
instrument	0.034532006	0.740000000	0.090464548	0.109727165
weather	0.025038925	0.641509434	0.083129584	0.100651273
night	0.022345203	0.611111111	0.080684597	0.097633136
flight, into	0.024171261	0.742857143	0.063569682	0.077797726
adverse	0.012484700	0.588235294	0.048899756	0.059880240
factors, accident	0.010149885	0.526315789	0.048899756	0.059737157
vfr	0.016872899	0.750000000	0.044009780	0.054216867
follow	0.013084684	0.642857143	0.044009780	0.054086538
spatial	0.015240445	0.789473684	0.036674817	0.045317221
maneuvering	0.007830619	0.535714286	0.036674817	0.045072115
failure, control, airplane	0.007830619	0.535714286	0.036674817	0.045072115
traffic	0.011684423	0.700000000	0.034229829	0.042270531
spin	0.009961638	0.636363636	0.034229829	0.042219542
stall, spin	0.010438911	0.684210526	0.031784841	0.039274924
pilots, flight	0.008811639	0.619047619	0.031784841	0.039227520
with, factors	0.007489462	0.565217391	0.031784841	0.039180229

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and, thus, are useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

In figure 8, the PCA shows the importance of the same terms, as well as the dangers of engine failure and loss of power. Terrain is demonstrated by the analysis to be a significant contributing factor within the region.

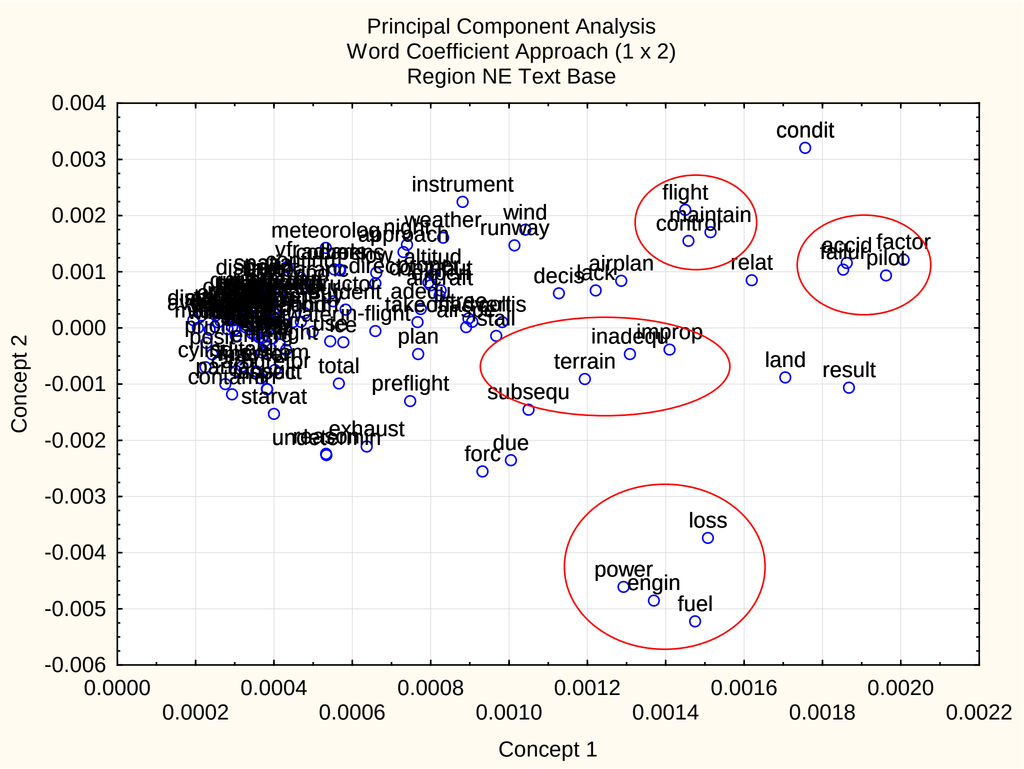

Figure 8. The PCA Coefficients (1 x 2), Region New England

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents and additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

The prototype text for the region is separated into three main clusters as per the analysis conditions. The first cluster is typified by failure to maintain adequate airspeed, resulting in a stall. The second cluster contains accidents relating to visual flight into IMC, particularly darkness. The third cluster consists of accidents wherein the pilot failed to maintain proper altitude in dark conditions. The previous report identified aircraft control, airspeed, VFR flight into IMC, and pilot judgment as the leading contributors to accident fatality within the region. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 8: SOUTHERN.

This section relates to the southern region, which consists of Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, Puerto Rico, South Carolina, and Tennessee. Fatal accidents within the region involve visual flight into IMC and stalling caused by various factors.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 9, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to stalling, airspeed, and loss of aircraft control, as well as adverse meteorological conditions, including the effects of these conditions, such as disorientation.

Table 9. The AIRES Results, Region Southern

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
into	0.036059842	0.714754098	0.104656745	0.126201227
altitude	0.022111094	0.560000000	0.100816131	0.120592627
inadvertent, stall	0.016644378	0.518005540	0.089774364	0.107557805
instrument	0.034777118	0.845744681	0.076332213	0.093309859
descent	0.013557506	0.509615385	0.076332213	0.091971310
low	0.015539070	0.565384615	0.070571291	0.085544693
weather	0.012135431	0.525490196	0.064330293	0.078024921
maneuvering	0.016025102	0.613207547	0.062409986	0.076076779
uncontrolled	0.013657234	0.583732057	0.058569371	0.071420208
maintain, while	0.011367328	0.553921569	0.054248680	0.066190253
night	0.014062461	0.623595506	0.053288526	0.065217391
continued	0.017202872	0.740740741	0.048007681	0.059052793
meteorological	0.019554230	0.842592593	0.043686990	0.053909953
disorientation	0.021390981	0.891089109	0.043206913	0.053361793
maintian, airspeed, while	0.012660769	0.661764706	0.043206913	0.053141238
airspeed, while	0.012493749	0.656934307	0.043206913	0.053134963
continued, flight	0.019371744	0.854368932	0.042246759	0.052163604
aircraft, control	0.009250519	0.567741935	0.042246759	0.051843997
known	0.007543712	0.517857143	0.041766683	0.051176471

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identified several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and, thus, are useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

In figure 9, the PCA shows the importance of the same terms, as well as dangers inherent to power loss and loss of control. Generally, the terms for this region that correlate highly to fatal accidents on a national level are indistinct from the majority of terms within the corpus. Further, a minority of accidents within the region involved nose gear collapse.

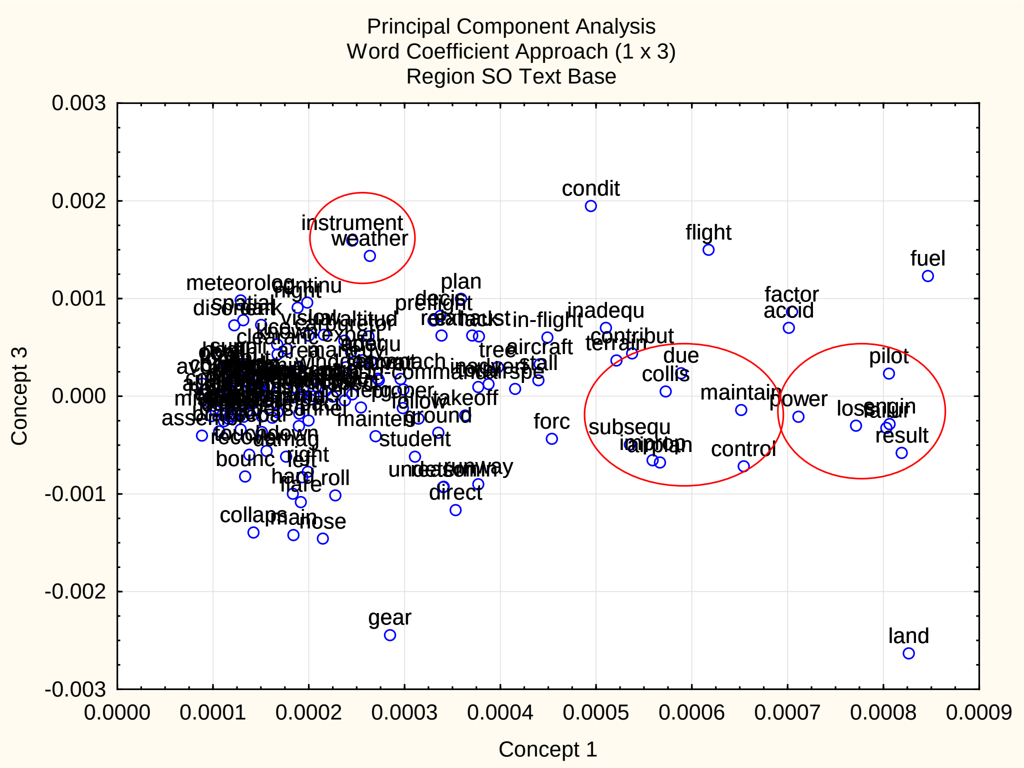

Figure 9. The PCA Coefficients (1 x 3), Region Southern

K-MEANS CLUSTERING. The data contain words and clusters of words that can be used to classify the documents and additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

The prototype text for the region is separated into three main clusters as per the analysis conditions. The first cluster within the region indicates that failure to maintain adequate airspeed, resulting in a stall, is often the cause of a fatal accident. The second cluster demonstrates the importance of avoiding visual flight into IMC, particularly fog and low cloud ceilings, as this could result in a fatality due to terrain collision. The third cluster again identifies failure to maintain adequate airspeed, resulting in a stall, but further identifies collision with terrain or trees as the cause of the fatalities. The separation of the first cluster from the third cluster likely relates to the specific choice by some NTSB agents to specifically call out the terrain collision in the report. The previous report identified airspeed, VFR flight into IMC, aircraft control, and in-flight planning/decisions as the leading contributors to accident fatality within the region. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicates that the different approaches achieve a similar result, though they are not directly comparable.

REGION 9: EASTERN.

This section relates to the eastern region, which consists of Delaware, Maryland, New Jersey, New York, Pennsylvania, Virginia, West Virginia, and Washington, D.C. Fatal accidents within the eastern region typically involve stalling due to either failure to maintain adequate airspeed or improper use of controls.

THE AIRES RESULTS. Information gain, precision, recall, and Weighted F-Measure are used in determining the significance of the pattern. Information gain represents to what extent each pattern predicts whether or not the accident will be fatal. This measure is bounded between 1 and 0, with 1 representing a perfect predictive gain and 0 representing a complete lack of predictive gain. An information gain of 1 would demonstrate that the inclusion of the corresponding pattern would completely predict all relevant outcomes, whereas an information gain of 0 would represent no increase in predictive accuracy. Precision is a measure of the ratio of retrieved accidents that are relevant, where a maximum value of 1 represents that each retrieved accident is relevant (fatal). A value of zero for this ratio would indicate that none of the accidents retrieved by using the pattern were relevant (fatal). Recall is the number of relevant (fatal) accidents that are retrieved, where a maximum value of 1 represents that every relevant (fatal) accident has been retrieved through using the pattern. A value of zero for this ratio would indicate that none of the relevant (fatal) accidents have been retrieved by using the pattern. A high precision implies that the returned pattern was relevant to the research question, while a high recall indicates that fewer accidents of the relevant (fatal) pattern have been missed. The F-Measure is the harmonic mean of precision and recall and is used to combine both measures into one figure for ease of reporting.

The AIRES results, as shown in table 10, highlight several terms and phrases that are highly correlated to fatal accidents. Among these are several terms relating to stalling, airspeed, and altitude as well as adverse meteorological conditions, including such effects of these conditions as spatial disorientation and collision with terrain.

Table 10. The AIRES Results, Region Eastern

Pattern	Information Gain	Precision	Recall	Weighted F-Measure
into	0.036785158	0.689655172	0.10989011	0.132100396
instrument	0.040670855	0.842105263	0.087912088	0.107095047
weather	0.019755254	0.567901235	0.084249084	0.101545254
failure, maintain, airspeed	0.014276943	0.502958580	0.077838828	0.093674234
low	0.014210271	0.527397260	0.070512821	0.085290208
night	0.012653337	0.526717557	0.063186813	0.076683707
meteorological	0.029103906	0.858974359	0.061355311	0.075348628
vfr	0.028782033	0.876712329	0.058608059	0.072055843
spin	0.017435334	0.765625000	0.044871795	0.055279783
disorientation	0.022216086	0.903846154	0.043040293	0.053167421
adverse	0.011722243	0.630136986	0.042124542	0.051790137
fog	0.014896967	0.754385965	0.039377289	0.048587571
spatial	0.023221076	0.976744186	0.038461538	0.047608252
imc	0.015788643	0.803921569	0.037545788	0.046390586
failure, airspeed	0.007391029	0.525641026	0.037545788	0.046108862
maneuvering	0.012630211	0.714285714	0.036630037	0.045207957
terrain, factors	0.013727940	0.800000000	0.032967033	0.040788579
maintain, altitude	0.008110451	0.590163934	0.032967033	0.040641228
collision, with, terrain	0.006189543	0.514285714	0.032967033	0.040558810

PRINCIPAL COMPONENTS. To verify the results of the AIRES analysis on a single-term basis, the PCA is conducted on the data to identify word clusters that cause variance within the corpus of documents. The PCA identifies several clusters of words, many of which are not discovered by the AIRES algorithm. Words and clusters that are higher on either axis represent greater variability within the documents and, thus, are useful as discriminators. Words and clusters that are both high on the axis and separate from the majority of words are both highly descriptive and unique, making them good classifiers of the documents.

In figure 10, the PCA shows the importance of the same terms, as well as the need for the pilot to maintain control in these unfavorable conditions and remain cautious during takeoff, especially as a student pilot.

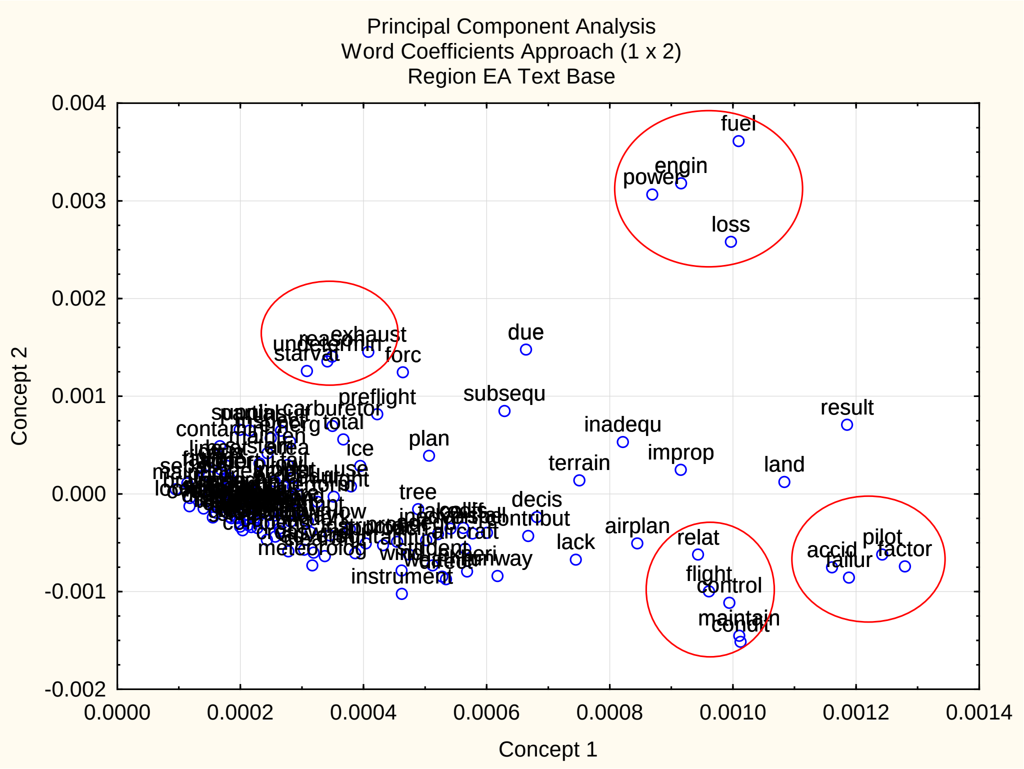

Figure 10. The PCA Coefficients (1 x 2), Region Eastern

K-MEANS CLUSTERS. The data contains words and clusters of words that can be used to classify the documents and additional techniques can be used to identify the most descriptive clusters. These clusters can then be examined for correlation to the desired criteria—in this case, accident fatality. The archetypical documents from each cluster not only provide patterns similar to the AIRES and PCA results, but they also provide context.

The prototype text for the region is separated into three main clusters as per the analysis conditions. The first cluster identifies the importance of maintaining airspeed to avoid stalling, while the second cluster indicates that stalling can also occur as a result of improper use of control inputs. The third cluster is not as well characterized as the others; however, it indicates adverse weather conditions, including darkness, low cloud ceilings, and failure to comply with instrument approach procedures. The previous report identified airspeed, VFR flight into IMC, aircraft control, and in-flight planning/decisions as the leading contributors to accident fatality within the region. The PCA, K-Means Clusters, and AIRES analysis share many of the same terms within their patterns, which indicate that the different approaches achieve a similar result, though they are not directly comparable.

CONCLUSIONS

Based on both an analysis of the national data as an aggregate and as a sum of individual analysis by region, there are several possible primary causes of fatal accidents within the nation. All regions included in this report experienced a significant number of fatal accidents as the result of pilots’ failure to maintain adequate airspeed, resulting in stalling. Additionally, aircraft operated under visual flight rules (VFR) flying into instrument meteorological conditions (IMC), especially darkness, fog, or low cloud ceilings, is frequently fatal. Instances of this behavior are both inadvertent and intentional and can be exacerbated by pilot overconfidence, lack of training, or lack of recent flights in these conditions. Although there are several minor effects associated with particular regions, they are not significant on a national scale. Chief among these secondary factors is pilot impairment due to alcohol or other intoxicants. While this is almost certainly a contributing factor to incidents, it lacks a strong positive correlation with fatality on a national scale. By contrast, the K-Means clustering process only detects engine failure (and subsequent loss of power) as a high fatality cluster on a national level despite loss of power being regularly identified in the Principal Components Analysis (PCA) of each region.

This report's analysis of the unstructured text portion of the National Transportation Safety Board (NTSB) accident reporting database is in general concurrence with the results of the logistic regression performed in the previous report (DOT/FAA/TC-12/52). That report's use of the structured text fields of the database (particularly the sections about leading causes) allows for much more specific conclusions to be drawn based on highly selective criteria. As a result, the logistic analysis loses some distinguishing characteristics of the data present only in the text portions. This loss is more than offset by the increase in accuracy and specificity provided by the logistic analysis and the degree of statistical assurance of the significance of the report's findings. Text analysis techniques are revealed to be useful if suited to exploratory and confirmatory analyses. For instance, the observed relationship between loss of engine power and fatality can be examined using a traditional regression analysis to determine its significance to fatal accidents. Likewise, the trends observed within the text data confirm the results of the traditional analysis and do not provide counter-indicated results, further testing the robustness of the initial analysis.

The Aviation Safety Information Analysis and Sharing (ASIAS) Information Retrieval and Extraction System (AIRES) software, developed by MITRE, appears to be successful in identifying keywords and phrases within the unstructured text of the NTSB reports, despite its unfinished state. In its present form, it provides superior results when compared to an unstructured (ranked) PCA in terms of isolating relevant terms as confirmed by both the K-Means clustering approach and traditional regression techniques. Without the ability for field experts to "train" the algorithm as it is intended, it will likely deliver nebulous results when compared to clustering techniques, particularly those that use intelligent algorithms. The clustering techniques provide, through prototype text, a more complete view of common input within a particular relationship. The intended functionality of the software would combine positive aspects of this view with algorithmic determinations, which will likely result in an overall superior analysis.

FUTURE RESEARCH.

STOP, PHRASE, AND SYNONYM LISTS FROM EXPERTS. One of the fundamental limitations of text analysis is the validity of terms automatically selected by the processing algorithm. With this type of analysis, terms such as “pilot” and “aircraft” are too common within the data set to represent the inclusion of useful information. Likewise, the presence of certain words is detrimental to the goal of the analysis because they represent post facto judgments that are of no use in generating predictive models. For example, the words “fatal,” “autopsy,” and “coroner” correlate highly with fatal accidents and are of no use in prediction. A related problem occurs because of the presence of synonyms within the data. A relevant group of synonymous words may individually be too low in frequency to merit inclusion, yet, as a group, reveal useful trends within the data. Common phrases also suffer from this effect, wherein the information gain of a specific-ordered group of words generates an information gain when used in a predictive measure. Further analysis could address these issues through consultation with field experts to create comprehensive lists of these relationships. Inclusion of these lists in the analysis would result in increased accuracy overall, and possibly the identification of alternate risk factors for fatal accidents.

NEURAL NETWORKS AND SELF ORGANIZING MAPS. K-Means clustering is a rather dated technique, having first been used in the late 1960s. Among the drawbacks of this approach is its computational unattractiveness—that it is classified as an NP-hard algorithm. This makes the processing of large data sets time-consuming or impossible, based upon the size of the set in question. This problem is exacerbated when dealing with unstructured text, as the word frequency and singular value decomposition processes generate large volumes of data. A possible solution to this problem is the incorporation of intelligent search methodology as implemented in a neural network. Using a neural network, such as a self-organizing map, to cluster the data provides the added benefit of mitigating potential overfitting because the process is nonexhaustive. Further research would include an implementation of neural network algorithms to replace K-Means clustering as a method of sorting the data based on common elements.

ANNOTATED BIBLIOGRAPHY

1. Bazargan, M. and V. Guzhva, 2007, “Factors Contributing to Fatalities in General Aviation Accidents,” *World Review of Intermodal Transportation Research*, Vol. 1, No. 2, pp. 170-182.

   The authors identify factors that increase the chance of a fatal GA accident by applying logistic regression on a large sample of GA accidents from 1983 to 2002. In this study, a wide range of variables, such as light condition, flight phase, pilot’s experience, pilot’s gender, wind condition, and aircraft characteristics and complexity involved in the accident, are analyzed.

2. Bazargan, M. and V. Guzhva, 2011, “Impact of Gender, Age and Experience of Pilots on General Aviation Accidents,” *Accident Analysis & Prevention,* Vol. 43, No. 3, pp. 962-970. DOI: 10.1016/j.aap.2010.11.023

The authors conduct a series of statistical analyses to investigate the significance of a pilot's gender, age, and experience in influencing the risk for pilot errors and fatalities in GA accidents. There is no evidence from the Chi-square tests and logistic regression models that support the likelihood of an accident caused by pilot error to be related to pilot gender. However, evidence is found that male pilots, those older than 60 years of age, and with more experience, are more likely to be involved in a fatal accident.

3. Bazargan, M., H. Kosalim, M. Williams, and A. Singh, 2012, "Comprehensive Analysis of General Aviation Accidents," DOT/FAA/AR-12/2.

   The authors conduct comprehensive analyses of the initiating causes for Fatal, Serious, and Minor/None GA accidents in all nine FAA regions. This report provides exploratory statistics based on month, time of day, phase, and purpose of flight. The report also explores the role of the top ten initiating causes for Fatal, Serious, and Minor/None GA accidents and examines associations between GA accidents and pilot experience. Furthermore, the report explores the role of aircraft complexity (based on engine horsepower) in Fatal, Serious, and Minor/None GA accidents.

4. Bazargan, M., H. Kosalim, and M. Williams, 2013, "Statistical Analyses For General Aviation Accidents," DOT/FAA/TC-12/52.

   The authors conduct a statistical analysis of the factors contributing to fatal general aviation (GA) accidents. The analyzed factors include flight rules, light conditions, weather condition, flight phases, pilot characteristics, and aircraft complexity. The study focuses on identifying associations and patterns between contributing flight elements and risk factors. The binary logistic regression analysis is conducted with the goal of predicting values of a categorical-dependent variable, and the dependent value in this study is the probability of an accident falling into a fatal category. This analysis statistically identifies the factors that increase the likelihood of a GA accident in each region becoming fatal.

5. Delen, D. and M.D. Crossland, 2008, "Seeding the Survey and Analysis of Research Literature With Text Mining," *Expert Systems with Applications*, pp. 1707-1720.

   The authors propose to enable a semi-automated analysis of large volumes of unstructured information through the application of text mining. This analysis allows the researcher to consider objectively all the information gathered and then apply the techniques of categorization to identify significant groupings of journals. A structured modeling method called IDEF is introduced in this paper.

6. Jeske, D.R., and R.Y. Liu, 2007, "Mining and Tracking Massive Text Data: Classification, Construction of Tracking Statistics, and Inference Under Misclassification," *Technometrics*, pp. 116-129.

   The authors identify that analyzing nonstandard data sets will be a challenging task. As the information contained in the data sets can be textual, image, functional data, or high-dimensional data without specified models, the analysis process is going to be complex.

Thus, the authors offer a systematic data-mining methodology that caters to a large free style text report database and helps to create a tracking statistic from the analysis. The proposed procedure is applied to the analysis of a Federal Aviation Administration aviation safety report project. The Program Tracking and Reporting Subsystem database of the FAA is used for the study. For the text classification method, the naïve Bayes classifier is adopted.

7. Kloptchenko, A., T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. Visa, 2004, "Combining Data and Text Mining Techniques for Analysing Financial Reports," *Intelligent Systems in Accounting, Finance and Management*, Vol. 12, No. 1, pp. 29-41.

The authors establish that the common practice of automatically mining financial figures from regulatory reports can be combined with analysis of the text portions to increase predictive accuracy. This improvement is centered around their research assertion that the textual component of annual reports contains "richer" information than financial ratios. The textual portion of the report can also be used to provide context for the ratio data. To evaluate and categorize the text, the authors implemented an intelligent search algorithm known as a "Self-Organizing Map" or "Kohonen Map." This technique is a form of neural network that applies intelligent search theory to a process similar to K-Means Clustering. Though the use of artificially intelligent algorithms was not within the scope of this report, it provides a platform for future work in the area.

8. Lee, S., J. Song, and Y. Kim, 2010, "An Empirical Comparison of Four Text Mining Methods," *The Journal of Computer Information Systems*, Vol. 51, No. 1, pp. 1-10.

The authors identify that the amount of textual data available for analysis is increasing at a dramatic rate within our society, necessitating the use of text analysis techniques. The paper examines four commonly used text mining techniques, identifying their characteristics and limitations. The examined techniques include latent semantic analysis, probabilistic latent semantic analysis, latent Dirichlet allocation, and the correlated topic model. The article is primarily intended to provide model selection guidance for parties seeking to analyze text data. The paper informs this report by codifying the terms subject matter experts use when describing text analyses, including corpus and the specific technical definitions of word and sentence.

9. Melby, P., 2011, "Mining Aviation Safety Reports Using Predictive Word Sequences," MITRE Corporation, Boston, Massachusetts.

The author describes a software program developed to solve the limitation the Aviation Safety Information and Sharing (ASIAS) faces, while reviewing the data sources that lack proper categorization. A text mining program is introduced by the author, called The ASIAS Retrieval and Extraction System (AIRES), which implements an algorithm to discover word sequences from aviation safety reports and makes a comparison of positively and negatively labeled records.

10. Salton, G., C. S. Yang, and C. T. Yu, 1975, "A Theory of Term Importance in Automatic Text Analysis," *Journal of the American Society for Information Science*, Vol. 26, No. 1, pp. 33-45.

The authors introduce a new technique, discrimination value analysis, as a basis for automatic indexing and content analysis. This technique ranks the text words according to how well they are able to discriminate the documents of a collection from each other. A strategy for automatic indexing is explained in detail by the authors.

11. Tseng, W., H. Nguyen, J. Liebowitz, and W. Agresti, 2005, "Distractions and Motor Vehicle Accidents: Data Mining Application on Fatality Analysis Reporting System (FARS) Data Files," *Industrial Management & Data Systems*, Vol. 105, No. 9, pp. 1188-1205.

The authors apply data mining techniques to fatal automotive accident reports provided by the National Highway Safety Administration. The research described in this article in many ways mirrors the research conducted in this report, though on a different topic. The authors used "Self-Organizing Maps" to categorize the accident data and determine which factors contributed to accident fatality. The results of the article suggest that an intelligent approach, such as the one employed by the authors, can be used successfully on accident data; this has implications for future work done on the National Transportation Safety Board database.

12. Wallace, B., A. Ross, and J.B. Davies, 2003, "Applied Hermeneutics and Qualitative Safety Data: The CIRAS Project," *Human Relations*, Vol. 56, No. 5, pp. 587-607.

The authors identify the two main approaches to text analysis—qualitative and quantitative—and discuss the strengths, weaknesses, and uses of both. They propose a system, the Applied Hermeneutic Methodology (AHM), which combines elements of both qualitative and quantitative analyses in an attempt to achieve a superior result. The AHM model is described and then applied to the Confidential Incident Reporting and Analysis System (CIRAS) used by railways in the United Kingdom. Though the new approach yielded some useful results, it has limitations primarily in the area of the representation of textual classifiers in numeric form. Further, the CIRAS data is confidential, making the use of positivist measures impossible.

13. Zhang, Q. and R.S. Segall, 2010, "Review of Data, Text and Web Mining Software," *Kybernetes*, Vol. 39, No. 4, pp. 625-655.

The authors conduct a review of a multitude of text mining and analytics software to provide a basis from which the software tools most applicable to a proposed analysis may be selected. The paper primarily consists of a discussion of the features, characteristics, and algorithms used by the selected software. This article was used in informing the decision about which software to select for comparison to the MITRE AIRES Software. Though the software package chosen for the report was not among those listed within the paper (STATISTICA), the discussion of algorithms was greatly informative to the process.

APPENDIX A—DETAILED METHODOLOGY

THE AIRES.

The negative consequences of aircraft accidents on manufacturers, operators, the industry as a whole, and the general public are well established. To promote the open exchange of safety information to facilitate continuous improvement in aviation safety, the Federal Aviation Administration has developed the Aviation Safety Information Analysis and Sharing (ASIAS) system. This system enables users to perform integrated queries across many different aviation safety databases. Models and insights developed using this system are then used throughout the industry to generate improvements in safety practices. In collaboration with the ASIAS initiative, the MITRE Corporation has begun work on a software program to solve the data sufficiency problems that exist in the structured fields of the ASIAS databases. The ASIAS Information Retrieval and Extraction System (AIRES) addresses the issue of structured field sufficiency by conducting an analysis of the more complete narrative report fields present within the databases. At a high level, the software functions by comparing positively and negatively labeled records to discover words and word sequences that have predictive power over a desired dependant variable. The identified words (and sequences of words) have a high precision and can then be used to classify reports that are related to the desired characteristic (dependant variable).

As demonstrated by the literature overview, there are many techniques for discerning predictive word sequences within a document. Most of these techniques rely on word frequency to reduce the dimensionality of the analysis, as a full analysis of a large sample is not computationally attractive. The developers of the AIRES software contend that aviation safety reports are characterized by highly precise word sequences, which are relatively rare. This view is consistent with the costly practice of having subject matter experts manually review accident reports for trends. The AIRES software program implements an approach that creates patterns based on their predictive power from the initiation of the analysis. Perhaps the greatest strength of the AIRES approach is its incorporation of what might be considered equivalent to a probabilistic latent semantic analysis with user training of the algorithm.

A document provided by Dr. Paul Melby of MITRE accompanies the AIRES Software, detailing the algorithm implemented by the AIRES software and providing a general problem description and suggestions for future work. In their initial analysis, the MITRE team identified the following features as being generally true of aviation safety reports:

- When an incident type or contributing factor is marked “true” in the raw data, there is nearly always support within the narrative as well.

- Most individual incident types or contributing factors are rare (only applicable to 1%-10% of the reports).

- The incident types and contributing factors are not exclusive (a report may discuss multiple incident types and contributing factors).

The proportion of contributing factors to overall incident types illustrates one of the strengths of the AIRES algorithm over frequency-based methods. The AIRES algorithm takes advantage of the particular characteristics of the aviation data in that even though there will be positive reports in the unlabeled data, the amount of unlabeled reports that are positive should be quite low. In a latent semantic analysis or principal component analysis, document frequencies and inverse document frequencies are used to reduce the data to more manageable levels. Therefore, one can see how the AIRES reduction technique will capture more predictive words than a frequency-based approach. When combined with a Porter stemming algorithm (new as of the last software revision), the dimensionality of the problem is reduced even further as words sharing the same root (ice, icing, iced) are reduced to their stemmed components and included as one word.

Another strength of the AIRES approach is its method of addressing word combinations when gaps exist between the words in a phrase. For example, in the statement "The operator proceeded to lose control of the aircraft before a collision with terrain," one could argue that the relevant words are "lose," "control," "collision," and "terrain." However, as they are separated by independent, uncorrelated words, a traditional analysis would include them only individually. The AIRES approach is capable of ignoring the interspersed words and constructing the phrase "lose control collision terrain," which may have better predictive power than any of those words individually. This is particularly true in the case of determining input to a fatal accident, as the context provided by "collision terrain" is much more likely to indicate fatality than individual uses of those words, which could also indicate "collision avian" for a bird strike or "visually obstructed by terrain" for a Visual Flight Rules pilot's approach in a mountainous region.

The AIRES software program is still under development. Although the algorithms are fully implemented, the user report training functionality is currently inoperable. For the purposes of this report, this situation is acceptable because the technique used for comparison is entirely statistical and does not involve user training of an algorithm. However, it is reasonable to expect that, when implemented, this functionality combined with the involvement of a field expert will produce significantly better results than the software alone.

STATISTICA.

MITRE's AIRES package is a highly automated high-level implementation of its underlying algorithm. A user with little technical knowledge can be quickly trained in the use of the software and begin applying domain knowledge almost immediately. By contrast, the other software packages reviewed for use in this report are considerably more manual and knowledge-intensive. Based on a review of the most commonly used text mining software, as identified in the Zhang and Segall article, the research team decided to conduct the analysis in the STATISTICA software package. This decision was based on the algorithms, the software implements, its quality and robustness, and its accessibility to academic users. The following nine-step process was used to reach conclusions about the causes of fatal general aviation accidents and to assess these conclusions against the previous regression analysis.

1. Data treatment: The fundamental application of text mining is to index the provided text in a fashion that meaningfully represents the number of times words appear within a document set, or corpus. The STATISTICA text mining package supports several parameter sets affecting how the importing of text data is processed. Due to the nature of text mining, there are several input parameters that can greatly affect the results of any subsequent analysis. For the purposes of this analysis, a word is defined as having consecutive characters with a minimum length after stemming of 2 and a maximum length of 25. The settings for the maximum number of consecutive vowels (4), consonants (5), duplicate characters (2), and punctuation (1) reduce the use of words that are not a part of the natural language. In the specific case of the National Transportation Safety Board (NTSB) accident reports, this tactic reduces or eliminates the inclusion of details such as aircraft numbers and airport codes. The characters analyzed are all English letters and standard numerals, including hyphens. To further eliminate words that do not provide context for the analysis, a stop list was employed based on standard English terminology. This list includes pronouns (in subject, object, and possessive forms) and was extended for this analysis with the inclusion of the following terms: "plane," "aircraft," "airplane," "accident," "report," "flight," and "fatal." Due to differences in NTSB investigators' regional dialects and their usage of the language, inclusion of these words could lead to false correlations not representative of underlying accident causes. Because the measurement of injury seriousness is one of the analytical goals, it is inappropriate to examine it with respect to input. As the reports are written after the fact with full knowledge of the parameters of the accident, any report including the word "fatal" would almost certainly correspond to a fatal classification.

2. Word frequencies: Once the data have been properly prescreened and stemmed, a word frequency matrix can be constructed. This matrix represents the highest frequency words present across the corpus. In contrast to the AIRES algorithm, this process may neglect words that, while mentioned in a minority of reports, have a high predictive power. The importance of a word within the documents is evaluated with word frequency and inverse document frequency, which evaluates words within the document and within the corpus. The STATISTICA software package automatically constructs a table that includes the word (or stem, or phrase), its count, the number of documents it appears in, and an example of a base word for a stemmed result. This data is summarized in a matrix that uses each selected word as a column header and the accident report index on the first row. A further table is created by this process, which contains the length of the document in words, the number of words within the document represented by the frequency set, and a list of these words. This will be used later in the analysis when examining prototype documents for a given set of words.

3. Transforming word frequencies and singular value decomposition: The word frequencies tabulated based on the inverse document frequency scheme provide an overview of which words occur frequently and discriminate between them within the corpus. To render this information into numeric format, singular value decomposition is performed upon the matrix of selected words by document, using the inverse document frequency transformation. This decomposition computes the number of components required to explain all variance in the data.

4. Scree plot: Once the decomposition has taken place, the number of values to be retained for further analysis must be considered. To this end, a scree plot was produced to graphically interpret the meaningful number of components for inclusion. The traditional practice for determining the useful cutoff point is to examine the plot and locate the "elbow," where the marginal return for extending the analysis is roughly equal to the marginal computational cost of further inputs. To the right of the elbow, most of the variance is explained by random noise in the data, or "scree." For the corpora included in this report, this relationship generally began to level off at the third component and captured most of the data by the ninth component.

5. Word coefficients and principal component analysis: Once the cutoff point for the analysis is determined, the relationships between the words within the corpus can be examined. To introduce some element of expert input that mirrors the AIRES process, the principal components analysis (PCA) is performed to graphically demonstrate clusters for consideration. Words that are higher on either axis scale represent more document variation, and clusters of words are related. Through interpretation of this graph, one can examine clusters of words and intuit what common phrases involving those words might include.

   There are multiple methods for arriving at the PCA. The data can either be produced via an eigenvalue decomposition of a covariance matrix or via a singular value decomposition of the data matrix itself. Under either method, most of the clusters will be the same—the differences are mostly related to the particularities of the noise distribution and graphical representation. The sensitivity of the word coefficients scatter plot makes it somewhat more efficient for manual analysis as the clusters are better separated and defined.

6. Document scores and K-Means Clustering: An alternative to visualizing the problem space in terms of word variance is to organize the data by document ID. This displays the data in terms of relationships between documents in the same semantic space, which can be useful for examining the relationships between similar accidents in terms of their input characteristics. This is useful for illustrating the similarity of cause factors among accidents that have similar results. Another method for analyzing words related to accidents of a particular category is performing an untrained cluster analysis based on document scores. This process works by using the top components (which explain much of the variance via the scree plot) as inputs. By using the document scores rather than inverse word frequencies, the analysis will ignore much of the noise in the data, focusing instead on the underlying dimensions identified by the singular value decomposition. This reduces the need to prescreen the data before analysis and results in better defined clusters. However, the resultant clusters may be less inclusive than those selected by a field expert.

7. Basic K-Means algorithm: The basic algorithm of K-Means Clustering partitions an input number of observations (*n*, in this case, represents related document groups) into a number of clusters, with each observation belonging to the cluster with the nearest mean value. Given that the set of document scores is (d_1, d_2,…,d_n), where each of the scores is the eigenvector derived from the singular value decomposition, the clustering algorithm

will attempt to partition the *n* observations into *k* sets where ($k \leq n$), the sets themselves are represented by $S = \{S_1, S_2, \ldots, S_k)$ in such a way as to minimize the variation within a cluster as represented by the within-cluster sum of squares.

Where μ*i* is the mean of points within the set S_i

$$\arg\min \sum_{i=1}^{k} \sum_{dj \in Si} \left\| xj - \mu i \right\|^2 \tag{A-1}$$

In this way, the clusters of like accident reports can be examined as a group based on prototype clusters for each category, which represent items very near the mean of the category. The K-means clustering algorithm is computationally unattractive (NP-hard) and, therefore, infeasible for very large data sets; however, it is sufficient for this application. This process is sometimes referred to as "ANOVA in reverse" in that an ANOVA significance test evaluates the group-to-group variability against the within group variability when testing the hypothesis that the means in the groups are independent. The K-Means algorithm moves items (documents in this case) into different groups to maximize the ANOVA significance.

ADVANCED K-MEANS ALGORITHM. Rather than defining the number of sets (*k*), the STATISTICA software package can logically iterate through the data and dynamically determine the appropriate number of clusters for a minimum discrimination characteristic. In the case of this analysis, the process was bounded between two and 25 clusters with a smallest percentage decrease of 1%. This process is analogous to cross-validation (the process itself is v-fold cross-validation) wherein the algorithm is prevented from over-fitting the data, and the result is the optimal number of clusters.

TOP STORIES (DISTANCE TO CLUSTER CENTROID). By initiating a cross-tabulation matrix of the count and frequency of clustered documents that fall into each injury category, a combination of variables and unique values is created. The examination of these frequencies assists in determining the relationships between the cross-tabulated variables. This practice identifies which clusters correlate most highly to each accident class. The most representative stories from each of these clusters (as measured by distance from the cluster mean) can then be accessed to provide insight into the highlighted (clustered) words that separate them as being highly correlated with fatality. An examination of these clusters is conducted to evaluate their predictive usefulness.